图书情报工作
第65卷 第23期 2021年12月

基于深度学习的网络科技信息情报价值计算方法研究*

■ 张敏[1,2,3,4] 刘欢[2,3] 丁良萍[2,3] 范青[5]

[1] 中国科学院武汉文献情报中心 武汉 430071 [2] 中国科学院文献情报中心 北京 100190

[3] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

[4] 科技大数据湖北省重点实验室 武汉 430071 [5] 华中师范大学国家文化产业研究中心 武汉 430079

摘 要：[目的/意义]针对当前科研人员无法从海量的网络科技信息中及时甄别有情报价值的情报内容的问题，建立一套综合性情报价值计算方法，从而对网络科技信息的情报价值进行计算判断，最终帮助科研人员快速而准确地发现有情报价值的网络科技信息。[方法/过程]综合考虑情报外部特征与文本语义内容特征，利用深度学习(预训练语言模型)BERT方法构建基于文本语义内容特征的情报价值计算模型，利用深度学习模型的预测输出完成打分，并结合基于情报外部特征的原始计算方法得到最终的综合评价得分。[结果/结论]实验结果显示，基于文本语义内容特征的情报价值计算模型可以对情报按照情报价值得分进行有效的星级区分，弥补了基于情报外部特征的原始计算模型中星级区分度差的问题，最终的综合评价结果表明本文提出的情报价值计算模型在实际应用中也能够很好地满足科研人员的需求。

关键词：网络科技信息 情报价值计算 文本语义内容 BERT

分类号：G250

DOI：10.13266/j.issn.0252-3116.2021.23.008

大数据时代，网络科技信息数量呈指数增长，对丰富的网络科技信息完成动态监测并支持战略决策分析日益成为情报机构的一项重要工作[1]。但网络科技信息海量化、多源化和复杂化的特点为科研人员及时发现高价值的情报信息带来了困难与挑战。因此如何从海量的网络科技信息中快速而准确地甄别出有情报价值的情报内容也成为情报学研究的重要方向。

网络科技信息情报价值计算是属于网络信息资源评价研究的一种。在邹益民的研究中，“情报价值”被定义为：“情报价值是指情报与用户发生联系时，情报所具有的对人有用的属性，是情报的客观属性与用户需求的耦合。”[2]本文中沿用这一概念的定义，依据此定义，网络科技信息情报价值计算就是通过某种计算方法来获取网络科技信息所具有的对人有用的属性值。可以发现，对于情报价值计算来说，最重要的两个因素就是情报本身的客观属性与用户需求，二者缺一不可。事实上，笔者通过对情报价值计算相关研究进行调研，发现当前网络科技信息情报价值计算方法主要也是从情报外部特征和用户行为特征这两个方面开展。前者主要从指标体系的构建入手，关注网络科技信息的外在特征，如信息来源、客观性、及时性等[3]，利用定性或定量的方式来完成评价。后者则从情报的关注对象入手，分析用户群体的类别特性[4]，从而结合不同的用户偏好来判断网络科技信息的情报价值。无论是基于指标体系分析外部特征还是基于行为特征分析用户偏好，当前在深入挖掘情报内容本身的语义信息方面还存在不足。随着自然语言处理等新技术的发展，对于文本内容的深层次语义挖掘的相关方法越来越成熟。文本内容是网络科技信息的客观存在形式，文本的语义特征对于网络科技信息的情报价值判断也具有重要参考意义。

自然语言处理技术的蓬勃发展得益于深度学习方法的出现。深度学习（deep learning）的概念由G. E. Hinton等[5]于2006年提出，作为一种基于无监督特征学习和特征层次结构的学习方法，深度学习通过模拟

* 本文系国家自然科学基金项目“基于CityGML的三维古建筑语义建模研究”(项目编号:41801295)和中国科学院文献情报能力建设专项项目“网络科技监测平台智能分析核心能力升级”(项目编号:Y9290906)研究成果之一。

作者简介：张敏，馆员，博士研究生，E-mail:zhangmin2012@mail.whlib.ac.cn；刘欢，博士研究生；丁良萍，博士研究生；范青，讲师，博士研究生。

收稿日期:2021-06-16 修回日期:2021-09-09 本文起止页码:70-78 本文责任编辑:易飞

人类大脑的神经网络进行分析学习，解决了很多复杂的模式识别难题。网络科技信息的情报分析研究面临海量数据的挑战，探讨深度学习技术在网络科技信息中的应用方法也是大有必要的。

笔者尝试基于深度学习方法中的 BERT 模型，对网络科技信息的文本语义内容特征进行情报价值评估计算，并结合传统的基于外部特征的情报价值计算方法，最终形成一套综合性情报价值计算方法。在能源领域科技信息监测平台中，对本文所提出的计算方法进行了实际应用，并对应用效果进行了评价。

1 情报价值计算方法相关研究

目前国内外学者关于网络科技信息情报价值计算方法主要可以分为基于情报外部特征的方法和基于用户行为特征的计算方法。

1.1 基于情报外部特征的计算方法

情报外部特征在这里主要是指网络科技信息在生产、展示以及传播等过程中附带的一些外在属性，例如信息来源、信息类型、发布时间、语言、长度等。最早的有关网络信息评价的指标是由 B. Richmond 提出的"10C 原则"[6]，包括内容（content）、可信度（credibility）、批判性思考（critical thinking）、版权（copyright）、引文（citation）、连贯性（continuity）、审查制度（censorship）、可连接性（connectivity）、可比性（comparability）和范围（context）。之后的研究者又在此基础上补充了信息来源、文本格式[7]、评论[8]、时效性[9]、原创性[10]等。随着网络信息技术的发展，基于网络链接分析技术的评价方法也出现，其中最受关注的是由 L. Page 等[11]提出的 Pagerank 算法，通过分析网页之间的超链接关系，来计算网页内容的重要性，网络信息关联的链接数越多则反映该信息的重要程度越高，这与情报学中的引用关系分析类似，其主要利用了网络信息的超链接这一外部特征。相似的方法还有 J. M. Kleinberg 提出的网页排序算法 HITS[12]等。在近几年的研究中，研究者更加注重指标体系构建的科学性与完备性，如赵玉遂等[13]应用德尔菲法，通过专家咨询的方式建立网络健康信息质量评价指标，最终明确了信息特性、媒体特性和发布特性 3 个一级指标以及信息准确性、页面设计和编辑的权威性等 15 个二级指标。邓胜利等[14]的研究从用户视角出发，通过用户调研的方式，构建了由内容和设计 2 个一级指标及 7 个二级指标、7 个三级指标组成的评价标准框架。

刘建华等[15]提出了情报来源、情报类型、情报主题对象、情报科技相关度和情报主题相关度等 5 个指标，并对相关指标进行细化形成 31 个二级指标。在这些指标中，既包含了情报资源的外部特征，也包含了部分情报内容特征，形成了对情报价值的综合评价方法。这一方法率先将情报外部特征进一步深入到主题，即情报的文本内容维度。笔者将这种基于文本内容特征的方法进一步深入，通过深度学习方法学习文本内容的上下文特征，将文本内容与情报价值关联起来。

1.2 基于用户行为特征的计算方法

情报服务人员对网络科技信息开展动态监测与分析，最终目的是服务用户，所提供的情报服务是否符合目标用户的信息需求决定了信息服务的效果与质量。因此，通过分析用户的信息行为特征，有针对性地提供情报服务也是情报工作人员的重点努力方向。张洋等[3]通过对网络科技信息资源评价的相关研究进行综述，提出"要树立以用户为中心的评价理念"。在对网络科技信息进行情报价值计算时，大量的研究者也结合用户的行为特征进行了探究。

早在 2000 年，赵继海[9]提出的 8 项评价指标中，就把用户（audience）作为一项单独的指标。对于用户行为特征的考虑更多地是体现在信息检索系统中的资源评价与排序中，例如 H. Karodiya 等[16]通过对检索系统的用户进行分类，在对检索结果进行排序时结合用户的分类得到不同类别用户的排序结果。S. L. Price 等[17]、M. Han 等[18]、L. Tamine-Lechan 等[19]的研究都通过探究用户的兴趣与偏好，尝试构建个性化的检索服务。在近几年的研究中，王晓丽等[20]也提出了网络信息资源评价指标构建的原则，其中的导向性原则就提出用户年龄、认知习惯以及文化程度导致的不同用户对网络信息的需求不同。王晰巍等[21]的研究中显示了不同网络社群用户在信息交互中的效果差异，进一步说明了不用用户群体特征对于网络信息利用价值的评判具有较大影响。

用户一般更加关注信息内容本身，有研究者通过构建基于二元分类的信息过滤模型，根据用户的偏好对信息进行分类过滤，从而提供更有情报价值的信息。例如 R. Bing[22]和 N. Vatani 等[23]的研究中都关注了对信息内容中的词的特征，通过分析词频、同义词等构建用户兴趣模型，将信息内容与用户偏好关联起来。笔者同样借鉴了这样一种基于信息过滤的思想，将网络科技信息文本的内容与用户的关注度关联起来，通过收集用户认为有情报价值的文本与无情报价值文本作

为训练集,从而构建信息过滤的二分类模型。

2　研究方法

目前,许多网络科技信息情报价值计算分析方法通常根据情报的外部特征,如情报来源的权威性、情报的类型等构建相应的指标,从而进行情报价值的判断。这些外部特征在一定程度上反映了情报的价值,如来源于政府部门的情报资源通常具有较高的价值,外部特征得分较高。但是这种方法并没有对情报的文本语义内容进行深入探究。针对这个缺陷,笔者提出融合文本语义内容特征的情报价值计算模型,着眼于情报的文本语义内容层面,在情报来源的权威性、情报的类型、情报中内容监测对象的重要程度、情报的科技相关度和情报的主题相关度 5 个外部特征维度的基础之上,增加情报文本语义内容维度。并且综合所有的评估指标得到最终情报价值计算的结果。融合文本语义内容特征的情报价值计算模型技术路线如图 1 所示:

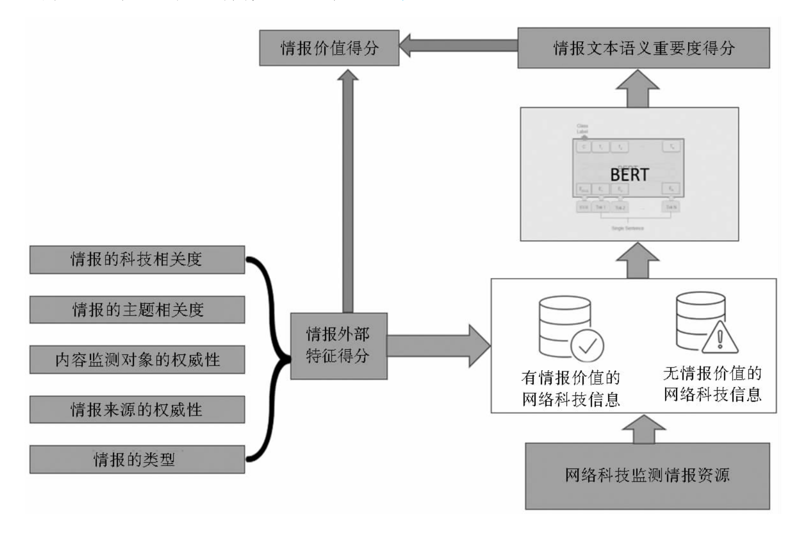


图 1　情报价值计算模型技术路线

笔者旨在充分利用预训练语言模型大规模无监督预训练和 Transformer 超强的文本语义、句法特征挖掘能力构建基于文本语义内容特征的情报价值计算模型。标注语料集是机器学习模型进行监督学习不可或缺的一部分,然而人工标注情报价值费时费力,因此笔者首先基于情报资源的外部特征得分自动构建情报价值计算的训练数据集,获取有情报价值的网络科技信息和无情报价值的网络科技信息。然后,笔者将情报价值计算定义为二分类任务,即预测网络科技信息为有情报价值或者无情报价值,通过模型对情报资源在有情报价值类别的预测置信度得到情报资源的文本语义重要度得分,最后综合文本语义重要度得分以及外部特征得分得到最终的情报价值得分。

2.1　情报价值计算语料构建

依托于笔者项目组开发的领域科技情报知识服务云平台,笔者构建了情报价值计算模型训练所需的语料集。领域科技情报知识服务云平台从情报工作的需求与工作流程出发,自动帮助情报人员从海量的网络科技信息资源中发现最新最重要的科技资源,借助信息抽取、自动分类、自动摘要、文本挖掘等方法,自动计算分析科技资源中包含的重要科技对象、重要科技术语,这些信息对构建情报价值计算模型的语料集有重要意义。

由于领域专家评估情报价值费时费力,而且不同专家之间可能存在意见分歧,构建大规模人工标注的情报价值计算语料集可行性不高。针对这个问题,笔者提出基于情报外部特征的语料集构建方法,充分利用情报的来源权威性、情报类型、主题相关度、监测对象的权威性、科技相关度这 5 个维度来自动构建情报价值计算的语料集。具体而言,这些外部特征是由情报分析人员根据经验知识制定的,笔者认为情报的外部特征一定程度上揭示了情报资源的重要程度,可设定重要度阈值来划分有情报价值的网络科技信息和无情报价值的网络科技信息,初步地自动构建一个大规模的情报资源计算数据集。

情报价值计算外部特征的框架如图 2 所示,基于上述的 5 个维度可以自动化地计算情报价值的外部特征得分。在领域科技情报知识服务云平台上搭建的能源领域科技信息监测平台采用基于外部特征得分的情报价值用于衡量情报的重要度,并反馈给用户。这种

模式运营多年,用户对于筛选排序的情报在一定程度上较为满意。因此,笔者选用平台中经过人工编译的报道以及外部特征重要度阈值≧0.6 的报道作为有情报价值的网络科技信息,没有经过人工编译的报道作为无情报价值的网络科技信息构建监督学习所需的数据集。这样可以间接将领域本体、领域主题词、热点词、科技主题词、重要监测对象等信息集成到模型中。

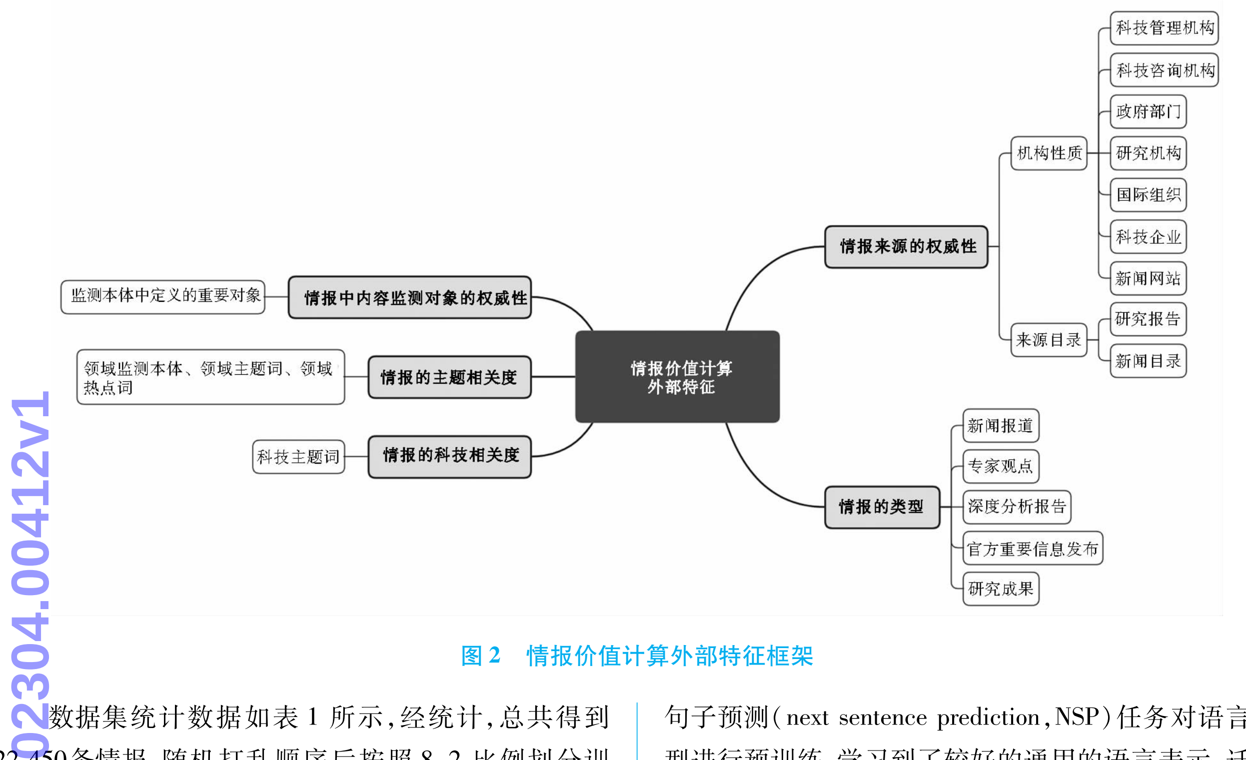


图 2　情报价值计算外部特征框架

数据集统计数据如表 1 所示,经统计,总共得到22 450条情报,随机打乱顺序后按照 8:2 比例划分训练集和测试集,训练集中有 17 959 条情报,测试集中共有4 491条情报。训练集中有情报价值的网络科技信息和无情报价值的网络科技信息的比值为 9 962:7 997,分布较为均衡。训练集和测试集中总计有情报价值的网络科技信息共 12 453 条,无情报价值的网络科技信息 9 997 条。笔者将有情报价值的网络科技信息标签赋予 1,无情报价值的网络科技信息标签赋予 0,构建二分类模型的初始语料。

表 1　数据集统计数据

数据集	无情报价值的网络科技信息/条	有情报价值的网络科技信息/条	合计/条
训练集	7 997	9 962	17 959
测试集	2 000	2 491	4 491
合计	9 997	12 453	22 450

2.2　模型架构

2018 年预训练语言模型 BERT[24] 的提出,引起了自然语言处理领域的广泛关注。许多研究者发现,在自然语言处理任务中使用预训练语言模型可以使下游模型性能得到较大的提升[25－26]。BERT 模型通过在维基百科等大规模无标注文本上使用 2 个预训练任务:掩藏语言模型(masked language model, MLM)和相邻句子预测(next sentence prediction,NSP)任务对语言模型进行预训练,学习到了较好的通用的语言表示,迁移到下游监督学习任务对于提升模型性能有很大程度的帮助。另外,Transformer 是一种超强的特征抽取器,通过自注意力机制,一定程度上解决了长短时神经网络的长距离依赖问题,能够对文本的语义、句法等特征进行很好的建模。笔者旨在充分利用 BERT 无监督预训练和 Transformer 模型架构的优势构建情报价值计算模型,同时将情报的外部资源特征融入到模型中辅助决策。笔者提出了基于文本语义内容的情报价值计算模型,其架构见图 3。

对于基于文本语义内容的情报价值计算模型而言,对于输入模型的情报资源首先进行文本向量化,将文本中的每个字映射到高维的向量空间中,获取字的表示。值得一提的是,在每句话之前添加[CLS]标识符,用该标识符的向量表示作为整句话的向量表示。然后输入到由 12 层 encoder 堆叠起来的 Transformer 模型中,获取[CLS]标识符的最终向量表示,输入到前馈神经网络并进行 SoftMax 分类,得到 BERT 模型对于无情报价值和有情报价值两个类别的置信度得分。使用 BERT 模型在有情报价值类别的预测得分作为情报的文本语义内容得分,与情报的外部特征得分按照 0.7:

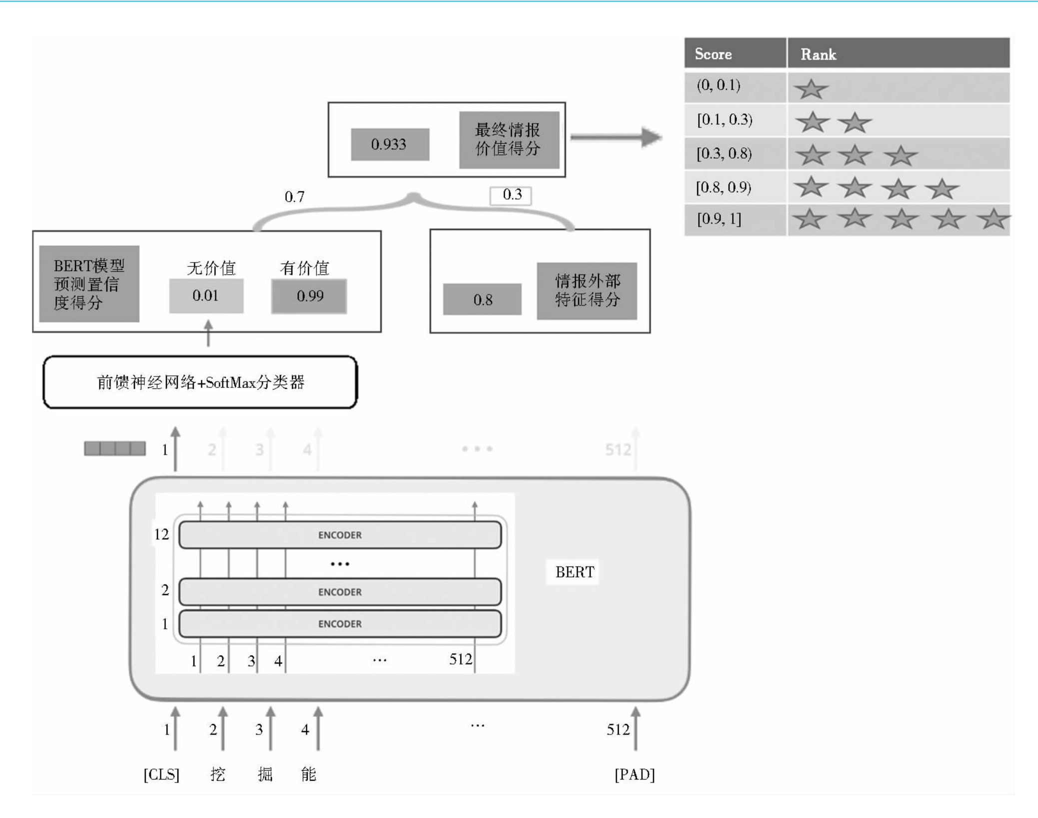


图3　基于文本语义内容的情报价值计算模型架构

0.3 的比重进行加权,得到最终的情报价值得分。根据该得分可以进一步设定阈值对情报重要度进行星级划分,当 0.9 ≦最终情报价值得分≦ 1 时,情报重要度为五星级;当 0.8 ≦最终情报价值得分 < 0.9 时,情报重要度为四星级;当 0.3 ≦最终情报价值得分 < 0.8 时,情报重要度为三星级;当 0.1 ≦最终情报价值得分 < 0.3 时,情报重要度为二星级;当 0 ≦最终情报价值得分 < 0.1 时,情报重要度为一星级。

3　实验及结果

笔者选定能源领域作为实验,构建情报价值计算模型。本节将介绍实验的具体操作步骤并对实验结果进行分析。

3.1　数据处理

笔者从能源领域科技信息监测平台获取到了能源领域的原始情报资源,如图 4 所示。由于情报资源是由半自动的方式自动收集整理得到,因此存在着大量的特殊标记、与正文无关的网页标题等噪声数据,这可能会影响对情报文本内容的语义分析。针对噪声问题,笔者首先对文本进行分句,然后通过对情报文本的分析,制定了一系列的规则来清洗其中的噪声。具体的规则示例如下:

```
crawlRecordId,filteredTitle,textCRLF
3298568.0,呼和浩特赛罕区加快锅炉煤改气,呼和浩特赛罕区加快锅炉煤改气 发布时间: 2014-08-19 来源: 中国环境报 大 中 小 内蒙古自治区呼
3298570.0,煤炭业出路在清洁高效利用,煤炭业出路在清洁高效利用 发布时间: 2014-08-19 来源: 经济日报 大 中 小 当前，煤炭行业遇到困境。
3298576.0,国家税务总局 国家能源局 关于落实煤炭资源税优惠政策若干事项的公告,国家税务总局 国家能源局关于落实煤炭资源税优惠政策若干事项的公告 发布时间:
3298579.0,四川省全面停征省级以下涉煤收费项目,四川省全面停征省级以下涉煤收费项目 发布时间: 2014-07-28 来源: 四川在线 大 中 小 近日
3298583.0,发改委部署煤炭业脱困 应对经济下行压力,发改委部署煤炭业脱困 应对经济下行压力 发布时间: 2015-05-14 来源: 经济参考报 大 中 小
3298587.0,山东煤监局加大30万吨以下关闭矿井监察力度,山东煤监局加大30万吨以下关闭矿井监察力度 发布时间: 2014-07-28 来源: 国家煤炭工业网 大 中 小
3298591.0,煤矿企业图纸监管加强,煤矿企业图纸监管加强 发布时间: 2014-08-18 来源: 经济日报 大 中 小 提高准确性和规范使用效果 国
3298593.0,河北: 主动关闭小煤矿最高奖励不低于350万元,河北: 主动关闭小煤矿最高奖励不低于350万元 发布时间: 2014-07-28 来源: 河北经济日报 大 中 小
3298598.0,我国将禁止销售和进口高灰分劣质煤,我国将禁止销售和进口高灰分劣质煤 发布时间: 2014-08-11 来源: 新华网 大 中 小 6 日从国家
3298603.0,黑龙江推进小煤矿整合 三年内煤炭企业总数控制在100家,黑龙江推进小煤矿整合 三年内煤炭企业总数控制在100家 发布时间: 2014-09-09 来源: 新华网
3298606.0,贵州拟定808处获准入煤矿名单,贵州拟定808处获准入煤矿名单 发布时间: 2014-08-11 来源: 中国煤炭报 大 中 小 8月1日，记者从贵
3298608.0,安监总局加快推进非煤矿山整顿关闭,安监总局加快推进非煤矿山整顿关闭 发布时间: 2014-08-15 来源: 新华网 大 中 小 新华网北京
3298610.0,贵州推进煤炭行业“万家企业”节能减排,贵州推进煤炭行业“万家企业”节能减排 发布时间: 2014-09-09 来源: 中国煤炭网 大 中 小 贵
3298612.0,国家能源局关于促进煤炭工业科学发展的指导意见,目录项的基本信息. 公开事项名称: . 国家能源局关于促进煤炭工业科学发展的指导意见 国能煤炭〔2015〕
3298615.0,三部门联合开展央企煤矿安全生产检查,三部门联合开展央企煤矿安全生产检查 发布时间: 2014-08-14 来源: 经济日报 大 中 小 国家
3298618.0,一季度煤炭产量同比下降3.5% 库存仍处高位,一季度煤炭产量同比下降3.5% 库存仍处高位 发布时间: 2015-04-16 来源: 人民日报 大 中 小
3298621.0,煤炭行业仍未“解冻” “颓势”或将持续,煤炭行业仍未“解冻” “颓势”或将持续 发布时间: 2015-04-16 来源: 新华网 大 中 小 新华能源
3298624.0,安监总局: 将关闭年产9万吨以下小煤矿,新华网北京7月24日电（记者徐博）国家安全监管总局局长杨栋梁在24日举行的50个煤矿安全重点县县委书记座谈会上说
3298630.0,不卖煤炭卖生态 福建资源大县永定经济转型的“富、美”追求,不卖煤炭卖生态 福建资源大县永定经济转型的“富、美”追求 发布时间: 2014-08-22 来源:
3298635.0,控煤成为河北今年治气头等大事 力争全年削减煤炭消费500万吨,控煤成为河北今年治气头等大事 力争全年削减煤炭消费500万吨 发布时间: 2015-04-21
```

图 4　能源情报网原始情报资源

(1)如果“加载更多:”或“参考资料:”或“原文出处:”或“推荐阅读:”或“责任编辑:”或“下一篇:”或“上一篇:”出现在句子中,则删除该句话;

(2)如果一句话以“来源:”或“编者按:”或“推荐CAJ 下载”或“PDF 下载.”或“HTML 阅读”或“下载频次”或“不支持迅雷”或“免费订阅”开头,则删除该句话;

(3)如果一句话中出现“发布时间”或“字号”或“来源:”,则将该句话中的“点击收藏”替换为空;

(4)如果句子长度<5,则删除该句话。

经过清洗之后,将外部特征得分≧0.6 的情报赋予标签1,得分<0.6 的情报赋予标签0,得到 BERT 模型训练的数据集,格式如图 5 所示。共得到训练集17 959条,测试集4 491 条。

0—→韩国开发氢燃料电池列车据国外网站报道， 韩国铁路研究所正在 参加 国家运输部的铁路技术研究项目 ， 将开发氢能铁路列车。该款列车是基于
1—→行业分析：互联网+电力改革，谁在风口上？随着电力体制改革的推进，原本垄断输配售三大环节的电网公司，将释放售电环节，并且重新厘定输配
2—→中电联召开第六届理事会第六次会议　　5月25日，中电联第六届理事会第六次会议在北京召开。杨昆常务副理事长报告了中电联一年来的工作情况
3—→“蓝光热井”池式低温供热堆或成供热新选择摘要：4月20日，河北省冬季清洁取暖典型案例展示交流活动在廊坊国际会议展览中心举行。《能源》
4—→到2040年，全球风电装机容量将增长15倍国际能源机构(IEA)周五表示，到2040年，海上风电业务将有望达到1万亿美元的规模，全球风电装机容量
5—→推广项目----中国科学院电工研究所。.推广项目.合作动态.推广项目.技术需求.所属公司...　推广项目..▪电工研究所科研成果汇编[2016-08-1
6—→波兰拟推能源新政 力争达到欧盟减碳目标据路透社报道，波兰能源部日前发布《2040年能源政策》提出，将不断降低对煤炭的依赖，以期到2030年
7—→财政部公布2016年钢铁煤炭去产能拟激励省份名单-新闻-能源资讯-中国能源网。今日，财政部网站发布《2016年钢铁煤炭去产能拟激励省份名单
8—→加拿大森科能源预计2020年石油产量或将增加5%加拿大森科能源公司周一预计，其2020年石油产量将增长5%，但表示，由于受到艾伯塔省削减产量
9—→亚洲洁能资本与雀巢共同打造最“绿色”光伏屋顶-新闻-能源资讯-中国能源网。根据此次签署的协议，亚洲洁能资本将为雀巢位于天津蓟县的生产基

图 5　清洗后的数据集格式

3.2　实验及结果

BERT 模型在 4 491 条测试集上测试的准确率为96.77%。笔者将 BERT 模型的情报价值预测得分与外部特征得分进行加权综合之后得到最终的情报预测得分,然后根据得分为情报划分星级。为了测试基于文本语义内容特征的情报价值计算模型的实验效果,笔者将其与基于外部特征的情报价值计算模型的效果进行对比。表 2 展示了两个模型的情报星级评价标准:

表 2　情报星级评价标准

星级	基于文本语义内容特征的情报价值计算模型评价阈值	基于外部特征的情报价值计算模型评价阈值
一星级	(0,0.1)	(0,0.6)
二星级	[0.1,0.3)	[0.6,0.77)
三星级	[0.3,0.8)	[0.77,0.82)
四星级	[0.8,0.9)	[0.82,0.88)
五星级	[0.9,1]	[0.88,1]

为了验证文本语义内容特征的融入对情报价值计算发挥了重要作用,笔者使用训练集和测试集的全部数据共计 22 450 条进行测试,统计各个星级下的情报数量分布。基于文本语义内容特征的情报价值计算模型和基于外部特征的情报价值计算模型的统计结果对比分析见图 6。

基于外部特征的情报价值计算模型的阈值划分是在多个领域下通过测试监测到的资源重要度分布比例而设定的。从图 6 可以看出,对于能源领域而言,此阈值对网络科技信息的情报价值预测得分基本上都集中在一星级和二星级,无法对网络科技信息的情报价值进行有效的区分,很难从大量信息中挑选出有情报价

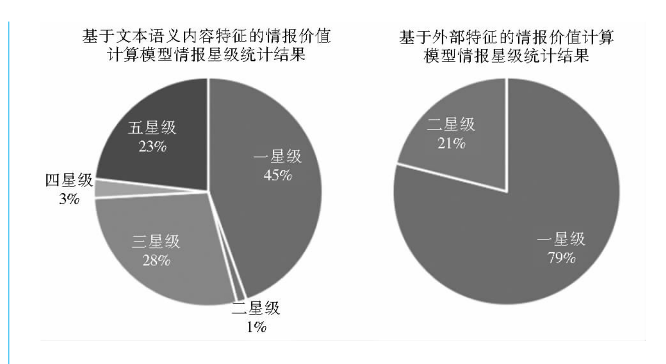


图 6　统计结果对比分析

值的网络科技信息来辅助情报人员分析决策,这也反映了基于外部特征的情报价值计算模型的领域局限性。而基于文本语义内容特征的情报价值计算模型可以对情报进行有效的星级区分,对有情报价值的网络科技信息和无情报价值的网络科技信息的区分度更高。

另外,笔者对其中一篇情报资源进行个例分析,样例如图 7 所示。针对这篇资源,原有的基于外部特征方法情报价值得分为 0,而 BERT 模型得到的情报价值预测得分为 0.999 978 423,综合外部特征和文本语义内容特征的模型的情报价值预测得分为 0.799 9。通过对情报文本进行分析,发现本文的方法可以将原本基于外部特征的情报价值计算模型预测为不重要的网络资源挖掘出来。笔者认为外部特征如情报来源的权威性虽然一定程度上可以反映网络科技信息的情报价值,但是有情报价值的网络科技信息仍然有可能潜藏在海量的网络资源中。只有融合了网络科技信息的文本语义特征,从文本内容本身出发,才能使情报价值的预测更具有可信性。

外媒：科学家发明可折叠石墨烯电池。外媒称，日前刊登在英国《自然·能源》杂志上的一项研究成果显示，科学家利用此前的发现创造出了一种能够存储能量的新型石墨烯折叠装置。据西班牙《世界报》网站2月17日报道，2004年，来自英国曼彻斯特大学的两名研究人员安德烈·海姆和康斯坦丁·诺沃肖洛夫，从石墨上剥离出非常薄的石墨烯碳层。事实证明，超薄的石墨烯具有柔韧性，是良好的热和电导体，比纸更轻，却比钢坚硬200倍。报道称，在第一阶段的研究和分类之后，围绕石墨烯的研究开始停滞不前。石墨烯的实际使用也举步维艰，这在很大程度上是因为其令人着迷的主要特性之一——超强硬度也使其很难加工。不过，近几年石墨烯再次成为焦点。2018年，美国麻省理工学院的西班牙科学家巴勃罗·哈里略-埃雷罗团队发现，当两层石墨烯以一个"神奇角度"缠扭在一起时，就会表现出非常规的超导电性。报道称，由此，一条全新的研究路线被打开了。如今，英国伦敦大学学院的科学家利用此前的发现创造出一种能够存储能量的新型石墨烯折叠装置。这种超级电容器最多可以180度对折而不损失性能，并且在经过5000次充电后仍可保持97.8%的电容。这种超级电容器的尺寸为6厘米，由两个电极组成，中间的胶片被用作传递电荷的介质。研究人员已经利用这种装置成功点亮了数十个LED灯。新成果解决了电池制造中反复出现的问题：难以在小空间中存储大能量。研究人员指出："我们采用了能使我们的超级电容器在具有高功率密度的同时又具有高能量密度的材料。通常情况下，只具备其中一个特征是可以实现的，但同时具备两个特征无疑是一个重大进展。

图 7　情报资源样例文本

4　应用效果评估

笔者选取了能源领域科技信息监测平台上监测的最近 500 条数据作为测试数据集，然后分别采用基于外部特征的情报价值计算方法和本文提出的综合性情报价值计算方法进行评分，最后按照各自情报星级划分标准得到对应的星级。情报的目的是被利用，满足用户的需要，解决问题。不同领域拥有其不同的需求与特征，什么样的网络科技信息更具有情报价值，应该由该领域的科研用户决定。领域内的专家用户与一般科研用户相比，专家用户对情报价值的判断更准确，评估水平更稳定。因此针对特定领域情报的价值应用效果评估，需要由该领域内的专家进行。由此，笔者邀请中国科学院武汉文献情报中心能源领域团队的 5 位专家对这 500 条数据进行星级评价，取平均星级。这 5 位专家都是能源领域科技信息监测平台的使用者，其中三位专家为长期从事先进能源科技情报研究的研究员，还有两位专家为具有博士背景的一线科研工作人员，他们对能够精准判断有情报价值的能源领域网络科技信息有着迫切的需求。评价标准采用完全认可、比较认可、比较不认可、完全不认可 4 个等级来表示通过两种计算方法得出的评价结果同领域专家评价结果的认可耦合度。其中，完全认可是指两者评价星级完全一致，用 0 表示；比较认可指两者评价的星级相差一个等级，用 1 表示；比较不认可是指两者评价的星级相差两个及以上等级，用 2 表示；完全不认可是指两者评价的星级相差三个及以上等级，用 3 表示。笔者通过上述两种计算方法得出的星级结果和专家评估的星级结果进行对比分析，其对比结果如表 3 所示：

表 3　评价星级结果对比分析

情报价值计算方法	完全认可/%	比较认可/%	比较不认可/%	完全不认可/%
基于外部特征	10	62	23	5
本文提出方法	20	67	11	2

如表 3 所示，基于外部特征的计算方法比较认可以上占 72%，而本文提出的计算方法比较认可以上占 87%，提高了 15%，因此，相比基于外部特征的情报计算方法，本文提出的综合性情报计算方法在实际应用中更能够使广大科研用户认可满意。同时该评估结果存在一定的不足，例如，专家组人数有限，代表性不够全面；不同专家的个人学识水平和主观需求不一样导致评估结果有倾向性；测试数据量不足导致可能的数据误差。

5　结语

笔者综合考虑情报外部特征与文本内容特征，利用深度学习 BERT 方法构建了基于文本语义内容特征的情报价值计算模型，从而对网络科技信息的情报价值进行判断，利用深度学习模型的预测输出完成打分，并结合基于情报外部特征的原始计算方法得到最终的综合评价得分。结果显示，基于文本语义内容特征的情报价值计算模型可以对情报进行有效的星级区分，弥补了基于情报外部特征的原始计算方法中星级区分度差的问题。与仅基于外部特征的原始计算方法相比，本文所提出的综合性情报计算方法能够更加有效地识别出有情报价值的网络科技信息，在实际应用中也能够很好地满足科研人员的需求。在后续研究中，将主要进行以下研究工作：

（1）情报价值计算语料的精炼。训练集的质量决

定了深度学习模型的实用效果,研究将针对训练语料,改进语料构建策略,完成精炼工作,根据模型的实际测试效果,循环迭代,形成更有区分度的有情报价值网络科技信息和无情报价值网络科技信息。

(2)扩充应用领域。研究将根据不同学科领域的语言特点与用户需求,尝试构建具有领域特点的情报价值计算模型。

参考文献:

[1] 张智雄，张晓林，刘建华，等. 网络科技信息结构化监测的思路和技术方法实现[J]. 中国图书馆学报，2014，40(4):4－15.

[2] 邹益民. 基于对象计算的情报价值判断方法研究[D]. 北京:中国科学院大学,2013.

[3] 张洋，张磊. 网络信息资源评价研究综述[J]. 中国图书馆学报，2010，36(5):75－89.

[4] 邹益民，张智雄. 网络科技信息情报价值评价方法综述[J]. 情报杂志，2014，33(5):25－30，59.

[5] HINTON G E，OSINDERO S，TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural computation，2006，18(7)：1527－1554.

[6] RICHMOND B. Cccccc. ccc (ten cs) for evaluating Internet resources[J]. Teacher librarian，1998，25(5)：20.

[7] STOKER D，COOKE A. Evaluation of networked information sources[C]//Proceedings of the 17th international essen symposium. Washington:ERIC，1994:287－312.

[8] SMITH A G. Criteria for evaluating information resources[J]. Public access computer systems review，1997，8 (3):1－14.

[9] 赵继海. Internet 信息评估:新世纪图书馆员的重要职责[J]. 大学图书馆学报,2000(5):35－38.

[10] 苏广利. 因特网信息资源评价研究[J]. 情报资料工作，2001(6):26－28.

[11] PAGE L，BRIN S，MOTWANI R，et al. The pagerank citation ranking：bringing order to the Web[R]. Stanford:Stanford InfoLab，1999.

[12] KLEINBERG J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM，1999，46 (5):604－632.

[13] 赵玉遂，许燕，吴青青，等. 应用德尔菲法构建网络健康信息质量评价指标体系[J]. 预防医学，2018，30(2):121－124.

[14] 邓胜利，赵海平. 用户视角下网络健康信息质量评价标准框架构建研究[J]. 图书情报工作，2017，61(21):30－39.

[15] 刘建华，张智雄. 情报重要度的指标体系和计算方法[R]. 北京:中国科学院文献情报中心，2011.

[16] KARODIYA H，SINGH A P D K. User specific search ranking technique[J]. International research journal of computer science engineering and applications，2013，2(1)：212－215.

[17] PRINCE S L，NIELSEN M L，DELCAMBRE L，et al. Using semantic components to search for domain-specific documents：an evaluation from the system perspective and the userperspective[J]. Information systems，2009，34(8)：724－752.

[18] HAN M，QIU X H. Personalized search engineer model[C]//Advanced materials research. Switzerland：Trans Tech Publications Ltd，2011：1216－1221.

[19] TAMINE-LECHANI L，BOUGHANEM M，ZEMIRLI N. Personalized document ranking：exploiting evidence from multiple user interests for profiling and retrieval[J]. Journal of digital information management，2008，6(5):354－366.

[20] 王晓丽，闫实，刘占波，等. 网络信息资源评价指标体系构建分析[J]. 软件，2020，41(5):53－56.

[21] 王晰巍，张长亮，韩雪雯，等. 信息生态视角下网络社群信息互动效果评价研究[J]. 情报理论与实践，2018，41(11):83－88,62.

[22] BING R. Information filtering algorithm based on feature vector [C]//Proceedings of the 2011 international conference on intelligence science and information engineering. New York：IEEE，2011：468－471.

[23] VATANI N，SHIRI M E. A personalized information filtering method[J]. International journal of computer science and security，2012，6(1)：1－8.

[24] DEVLIN J，CHANG M W，LEE K，et al. Bert：pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2020－12－25]. https://arxiv.org/abs/1810.04805.

[25] BELTAGY I，LO K，COHAN A. Scibert：a pretrained language model for scientific text[EB/OL]. [2020－12－30]. https://arxiv.org/abs/1903.10676.

[26] LEE J，YOON W，KIM S，et al. Biobert：a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics，2020，36(4)：1234－1240.

作者贡献说明:

张敏:方案设计、论文撰写及定稿;

刘欢:实验验证及论文校对;

丁良萍:数据整理及论文校对;

范青:方法应用及评价。

Research on the Web Technology Information Value Calculation Method Based on Deep Learning

Zhang Min[1,2,3,4] Liu Huan[2,3] Ding Liangping[2,3] Fan Qing[5]

[1] Wuhan Library, Chinese Academy of Sciences, Wuhan 430071

[2] National Science Library, Chinese Academy of Sciences, Beijing 100190

[3] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

[4] Hubei Key Laboratory of Big Data in Science and Technology, Wuhan 430071

[5] National Cultural Industry Research Center, Central China Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] In view of the problem that it's difficult for researchers to find valuable information from large amounts of scientific and technological information in the Web, this paper constructs a comprehensive calculation method for information value. It can calculate the information value of Web technology information and help researchers find Web technology information of information value quickly and accurately. [**Method/process**] Taking overall consideration of the external feature and textual semantic feature of the information, this paper used deep learning (pretrained language model) BERT to construct information value calculation model based on the textual semantic feature, used the predictive output of the deep learning model to complete the scoring, and combined the original calculation method of the external feature of the information to get the final information value score. [**Result/conclusion**] The experimental results show that the information value calculation model based on the textual semantic feature can rank the information to different levels according to their information value score, which makes up for the problem of poor star differentiation in the original calculation method only based on the external feature of the information. And the final comprehensive evaluation results show that the information value calculation model proposed in this paper can also meet the needs of researchers in the practical application.

Keywords: Web technology information information value calculation textual semantics BERT